УДК 004.8

МЕТОД ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ ТЕХНІК ПРОПАГАНДИ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

М. Молчанова ORCID: 0000-0001-9810-936X
Хмельницький національний університет, Україна
E-mail: m.o.molchanova@gmail.com

Анотація. *Робота присвячена створенню та апробації методу нейромережевого виявлення техніки пропаганди за маркерами з візуальною аналітикою, що дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та моделей керованого машинного навчання у вихідні дані, що містять числові оцінки присутності кожного пропагандистського прийому та розміченого тексту з візуальною аналітикою присутності виявлених пропагандистських маркерів. Було проведено дослідження, яке дозволяє виявити 17 основних пропагандистських прийомів. У дослідженні порівнювалися 3 підходи, які найчастіше використовуються: традиційний підхід машинного навчання, підхід на основі рекурентних нейронних мереж і підхід на основі трансформаторних моделей. Найвищих результатів досяг підхід на основі моделі трансформатора, який використовує механізми самоуважності, які дозволяють кожному елементу послідовності безпосередньо взаємодіяти з усіма іншими елементами. Це забезпечує ефективне захоплення довготривалих залежностей, що характерно для пропагандистських технік. Такий підхід дозволив виявити методи пропаганди з точністю до 0,96.*

Ключові слова: *BERT, RNN, методи пропаганди, виявлення пропаганди, пропагандистські маркери, візуальна аналітика*

1.Вступ

Пропаганда, замаскована під звичайні новини, поширюється вже багато десятиліть, але сучасна цифрова ера додатково створює умови для її більш швидкого, масового та ефективного поширення [1].

Створюються нові методи генерації текстів, які дедалі частіше мало відрізняються від створених людиною [2], що призводить до стрімкого зростання кількості контенту. Тому це все підкреслює важливість створення автоматизованих методів для виявлення пропагандистських маніпуляцій, які допоможуть користувачам отримувати інформацію більш усвідомлено.

Метою дослідження є підвищення точності виявлення методів пропаганди шляхом розробки методу виявлення методів пропаганди за маркерами на основі набору моделей машинного навчання, окремо для кожного методу пропаганди, навчених на модифікованих маркованих даних.

Основні внески дослідження можна підсумувати так:

– Розроблено підхід до підготовки навчальних даних, що дозволяє проводити навчання моделей машинного навчання для окремих технік пропаганди;

– Запропоновано метод виявлення прийомів пропаганди, що дозволяє знаходити силу прояву кожного з 17-и прийомів пропаганди, а також візуально інтерпретовувати отриманий результат з використанням моделі LIME.

– Експериментально продемонстровано ефективність використання нейромережевих моделей-трансформерів в порівнянні з рекурентними моделями та Traditional machine learning approaches

В цій роботі, далі представлено огляд пов'язаних робіт у сфері виявлення прийомів пропаганди згідно двох складових дослідження, зокрема аналіз наявних шляхів до вирішення проблеми виявлення пропаганди та аналіз моделей машинного навчання для виявлення прийомів пропаганди. Третій розділ паперу містить схему та кроки методу нейромережевого виявлення технік пропаганди за маркерами. Четвертий розділ присвячений опису плану експерименту виявлення технік пропаганди за маркерами та підготовці датасету. П'ятий розділ містить результати експерименту, їх аналітику та обговорення.

2. Аналіз останніх досліджень та публікацій

Наявні шляхи вирішення проблеми виявлення пропаганди. Проблема виявлення пропаганди залишається актуальною, адже і досі з'являються нові способи впливу на користувачів для поширення пропагандиських повідомлень. З огляду на це, виникає потреба у постійному моніторингу нових способів створення пропагандиського вмісту та вдосконалення методів їх ідентифікації, що є важливою задачею забезпечення інформаційної безпеки та протидії дезінформації. Тому науковці працюють над виявленням нових маркерів та нових прийомів пропаганди, а також над покращенням існуючих підходів для її виявлення.

Досліджено основні методи аналізу газетних текстів для виявлення маніпулятивних технологій, що допомагає застерегти від дезінформації та пропаганди [1]. Представлено новий набір еталонних даних чеською мовою для навчання та оцінки сучасних і майбутніх методів розпізнавання 18 маніпулятивних прийомів, таких як нагнітання страху, релятивізація та навішування ярликів. Показано, що поєднання контент-аналізу з запропонованим стильовим аналізом підвищує точність виявлення 15 з 17 оцінених маніпулятивних технік від 0.05% до 1.46%. Метод перевірено на пропагандистській базі QCRI. Подальші дослідження будуть зосереджені на додаванні нових стилометричних характеристик, вдосконаленні існуючих методів та використанні методів доповнення даних для боротьби з дисбалансом етикеток. Також планується перейти до дрібнозернистої класифікації на рівні проміжків часу, а не на рівні документа в цілому.

У ще одному дослідженні представлено багатомовний набір даних про пропаганду та проведено експеримент для дослідження маркерів, за якими людські анотатори та алгоритми класифікації відрізняють пропагандистські статті від непропагандистських на певну тему [3]. Показано, що перебільшення, зменшення описовості та відсутність адекватних джерел часто зустрічаються у пропагандистській пресі. Аналізатор VAGO підтвердив, що використання невизначених маркерів значно корелює з цими особливостями. Виявлено, що моделі машинного навчання ефективні для виявлення пропаганди на певну тему, але потребують покращення щодо пояснюваності та узагальнення на інші теми. Подальші роботи зосередяться на вдосконаленні аналізу, розробці багатомовних моделей та покращенні інструментів пояснюваності. Також планується введення нових міток для уточнення анотацій та ідентифікації більшої кількості стилістичних особливостей.

Застосування моделі MVPROP, що використовує багатовимірні контекстні вбудовування, дозволяє покращити точність виявлення пропаганди. Експерименти показали, що модель може бути перенесена на новинні статті[4]. Для тестування представлено TWEETSPIN, набір даних із твітами, що містить слабкі анотації тонких пропагандистських технік, і модель MVPROP для їх виявлення. TWEETSPIN включає лише ідентифікатори твітів, що відповідає умовам використання Twitter, і містить потенційно образливі та ворожі висловлювання. Основним обмеженням є слабкі анотації через великий масштаб даних. У майбутньому планується дослідження виявлення пропаганди на рівні окремих фрагментів.

Дослідники застосували мовну модель RoBERTa для виявлення пропагандистських технік у новинних статтях [5]. Модель оцінювалась за допомогою референсного набору даних для завдання SemEval-2020 Task 11, демонструючи здатність виявляти складні техніки пропаганди і перевершуючи базову модель з показником F1-score у 60.2%. У той час, як [6] аналізувались можливості використання великих мовних моделей (LLMs), зокрема моделі GPT-3.5-Turbo від OpenAI, для виявлення ознак пропаганди в новинних статтях. Використовуючи технологію, що лежить в основі ChatGPT, дослідники аналізували тексти для визначення присутності різних технік пропаганди, визначених у попередній роботі [7]. Розроблено ретельно уточнений запит, який поєднується зі статтями з мережі Russia Today (RT) та датасету SemEval-2020 Task 11, щоб визначити наявність пропагандистських методик. Дослідження показало, що технологія LLM може давати розумні висновки про пропаганду, хоча точність виявлення складає всього 25.12% за датасетом SemEval-2022. Однак вона демонструє потенціал як інструмент для виявлення пропаганди для кінцевих користувачів, таких як медіа споживачі та журналісти.

Як підтверджено у роботах вище, пропаганда характеризується прийомами, за які відповідають певні маркери, які притаманні використовуваним прийомам. У роботі увага буде зосереджена на виявленні 17 known прийомів пропаганди, детально описаних в [8]. А саме: «Appeal to fear-prejudice», «Causal Oversimplification», «Doubt», «Exaggeration», «Flag-Waving», «Labeling»,

«Loaded Language», «Minimisation», «Name Calling», «Repetition», «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches», «Whataboutism».

Моделі машинного навчання для виявлення прийомів пропаганди. У рамках дослідження будуть використані 3 підходи щодо моделей машинного навчання:

– Традиційний підхід на основі машинного навчання.

– Підхід на основі рекурентних рейромереж.

– Підхід на основі моделей-трансформерів.

Традиційний підхід на основі машинного навчання охоплює декілька методів і алгоритмів, призначених для розв'язання різноманітних завдань прогнозування, класифікації та кластеризації даних, в тому числі і для виявлення прийомів пропаганди. Лінійна регресія використовується для моделювання лінійних залежностей між вхідними функціями (ознаками) і цільовими значеннями, а також є одним з найпростіших методів регресійного аналізу і часто використовується для прогнозування числових значень. Метод опорних векторів (SVM) шукає оптимальну гіперплощину, яка найкращим чином розділяє два класи точок даних у просторі ознак. Він часто використовується для задач класифікації, особливо коли дані мають складну структуру [9]. Навчання на основі байєсових мереж ґрунтується на байєсівській ймовірнісній моделі, де кожна змінна розглядається як випадкова, і використовуються правила байєсівського висновку для побудови моделі. Також серед традиційні підходів машинного навчання для виявлення прийомів пропаганди використовується логістична регресія, k-NN, алгоритми кластеризації, наприклад, k-means. У роботі [10] автори наводять результати дослідження кілька моделей класифікації, включаючи мультиноміальний наївний метод Байєса, SVM, логістичну регресію та K-найближчих сусідів. У статті [11] автори представили два альтернативні методи (BERT та SVM) автоматичного визначення прокремлівської пропаганди в газетних статтях і дописах у Telegram.

Підхід до виявлення пропаганди на основі рекурентних нейронних мереж використовується для аналізу послідовних даних, зокрема текстів, що часто зустрічаються у соціальних мережах. Рекурентні нейронні мережі, Long Short-Term Memory (LSTM) і Gated Recurrent Unit (GRU) це різновиди архітектур рекурентних нейронних мереж, кожна з яких має свої особливості і застосування у вирішенні різних завдань у машинному навчанні, зокрема виявленні пропаганди в соціальних мережах. RNN є базовою архітектурою, яка здатна обробляти послідовні дані, зберігаючи інформацію у вигляді внутрішнього стану (пам'яті), який оновлюється при кожному новому вході [12]. LSTM – це розширена версія RNN, яка включає додаткові механізми, такі як ворота забування, ворота оновлення та ворота виходу. GRU являється спрощеною версією LSTM, яка має менше внутрішніх компонент. GRU вважається менш обчислювально витратною архітектурою порівняно з LSTM [13]. Результати дослідження ідентифікації пропаганди на платформі Twitter

під час пандемії COVID-19 авторів [14] показали, що запропонована пропагандистська ідентифікація на основі LSTM показала кращі результати, ніж інші розглянуті у роботі методи машинного навчання. За допомогою запропонованого підходу на основі LSTM досягається точність 77,15%. А у статті [15] автори використовують техніки глибокого навчання Bi-LSTM і Bi-GRU з методами SVM із слабким контролем забезпечили. Даний підхід забезпечив точність 90% у виявленні пропагандиських новин. Автори стверджують, що такий підхід є дуже ефективним і дієвим для нерозмічених даних.

Підхід на основі моделей-трансформерів передбачає використання таких архітектур нейромереж як BERT, RoBERTa, DistilBERT, GPT тощо [16]. BERT є однією з найвідоміших архітектур трансформерів, розроблена Google. BERT здатний досягати вражаючих результатів у завданнях обробки природної мови (NLP) завдяки здатності до контекстної обробки слів і здібності до підготовки звичайних моделей для багатьох NLP-завдань [17]. RoBERTa – це оптимізований підхід до BERT, який покращує навчання і результати моделі на різних NLP-завданнях шляхом застосування різних оптимізаційних стратегій [18]. DistilBERT вважається легковаговою версією BERT, яка зберігає сутність оригінальної моделі, зменшуючи кількість параметрів і зберігаючи високу продуктивність на різних NLP-завданнях. GPT – це родина моделей трансформерів, розроблених OpenAI. Даний підхід застосовують у дослідженні для класифікації пропаганди [19]. Автори використовують три моделі глибокого навчання, CNN, LSTM, Bi-LSTM і чотири моделі на основі трансформаторів, а саме багатомовні BERT, Distil-BERT, Hindi-BERT і Hindi-TPU-Electra. Експериментальні результати вказують на те, що багатомовні моделі BERT і Hindi-BERT забезпечують найкращу продуктивність із найвищим показником F1 84% за даними проведеного експерементального дослідження. Також у дослідженні [20] досліджується продуктивність BERT і RoBERTa, DeBERTa з комбінацією різних методів збільшення даних для виявлення пропагандистських текстів. Автори змогли досягти F1 micro 60% на тестовому наборі, використовуючи ансамбль моделей BERT, RoBERTa та DeBERTa.

Отже, наведені підходи знаходять своє застосування у завданні виявлення прийомів пропаганди.

3. Метод виявлення та класифікації технік пропаганди

Для реалізації методу виявлення та класифікації технік пропаганди за маркерами пропонується створити 17 моделей машинного навчання, кожна з яких буде відповідати за визначений прийом пропаганди. Такий підхід (рисунок 1) дозволить навчити моделі машинного навчання таким чином, щоб у них змогли вибудуватись залежності, притаманні конкретним видам пропаганди.

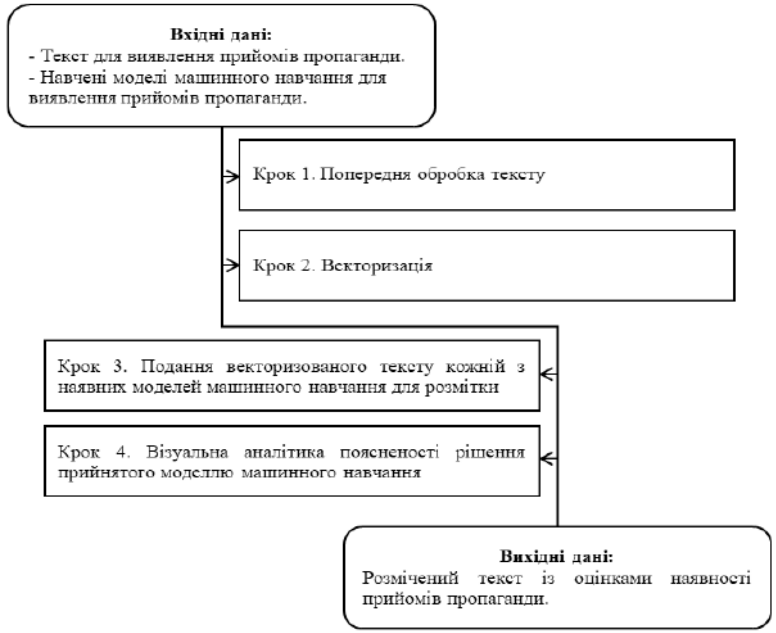

Рисунок 1. *Кроки методу виявлення та класифікації технік пропаганди*

В загальному, схема методу для виявлення прийомів пропаганди наведена на рисунку 1. Метод дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та навчених моделей машинного навчання у вихідні дані, які містять числові оцінки наявності кожного з прийомів пропаганди та розмічений текст з візуальною аналітикою присутності детектованих маркерів пропаганди.

Вхідними даними методу виявлення прийомів пропаганди є текст для виявлення прийомів пропаганди та навчені моделі машинного навчання для виявлення прийомів пропаганди.

Попередня обробка тексту включає видалення розділових знаків та стоп-слів, хоча і розділові знаки, розміщені певним чином, також можуть впливати на наявність пропаганди [22]. Асоціація споріднених слів була виконана шляхом лематизації, яка показує кращі результати, ніж стемінг. Для лематизації була використана відповідна стандартна бібліотека Pyton. Однак, в рамках даного дослідження такий вплив досліджуватись не буде.

Наступним кроком є векторизація тексту після попередньої обробки. Векторизоване представлення подається на вхід кожній навченій моделі машинного навчання, яка виконує передбачення наявності для кожного прийому пропаганди та його сили прояву. Більш детально крок 3 наведено на рисунку 2.

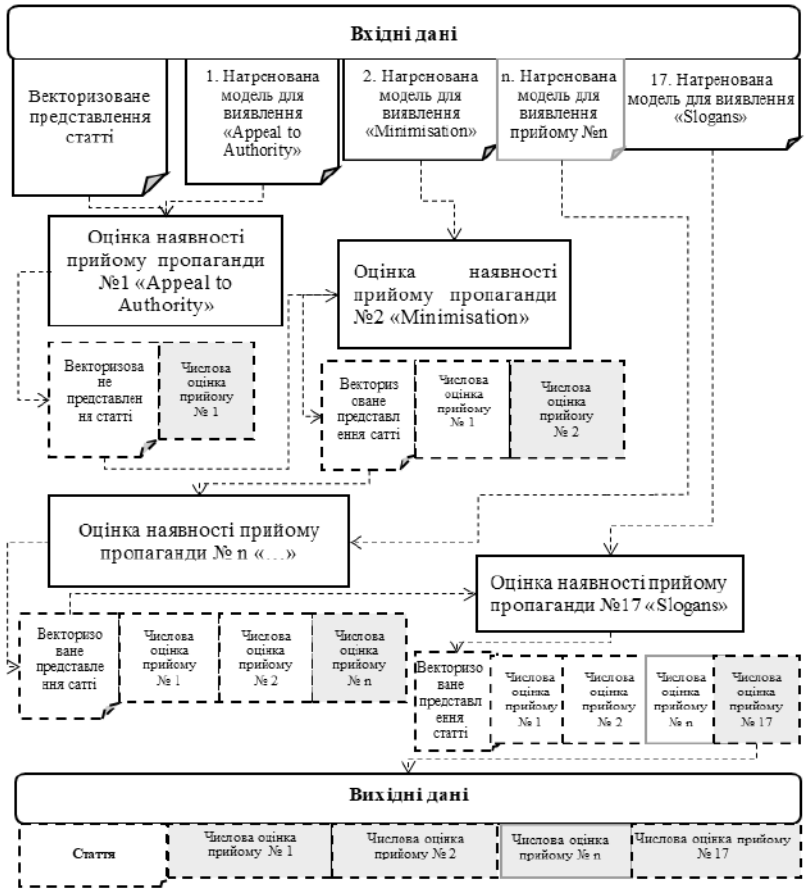

Рисунок 2. *Деталізація кроку подання векторизованого тексту кожній з моделей машинного навчання для розмітки*

Останнім етапом є проведення візуальної аналітики для поясненності рішення прийнятого кожною моделлю машинного навчання. Візуальна аналітика використовується із застосуванням методу Лайма, що є методом для інтерпретації прогнозів моделей машинного навчання, що розроблений для того, щоб пояснювати індивідуальні прогнози складних моделей [23]. LIME наближає будь-яку модель машинного навчання чорної скриньки до локальної, інтерпретованої моделі для пояснення кожного окремого прогнозу. LIME дозволяє зрозуміти, які частини вхідних даних вплинули на рішення моделі.

Вхідними даними кроку подання векторизованого тексту кожній з наявних моделей машинного навчання для розмітки є векторизоване представлення

статті та натреновані 17 моделей машинного навчання. Моделі по черзі оцінюють векторизоване представлення текстового контенту для аналізу на предмет наявності кожного з 17-и прийомів пропаганди. Вихідними даними є числові оцінки сили проявів прийомів пропаганди, що притаманні поданому векторному представленню тексту.

Таким чином, було створено метод нейромережевого виявлення технік пропаганди за маркерами, що дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та навчених моделей машинного навчання у вихідні дані, які містять числові оцінки наявності кожного з прийомів пропаганди та розмічений текст з visual analytic присутності детектованих маркерів пропаганди.

4. Експеримент

У ході проведення експерименту досліджувалось використання 3-х підходів до навчання моделей для виявлення прийомів пропаганди:

– Традиційний підхід на основі машинного навчання.

– Підхід на основі рекурентних рейромереж.

– Підхід на основі моделей-трансформерів.

У рамках традиційного підходу на основі машинного навчання досліджувалось виявлення прийомів пропаганди з використанням моделей регресії, SVM, Random Forest та Naive Bayes. Для прийомів «Appeal to Authority», «Black and White Fallacy», «Reductio ad Hitlerum», «Red Herring», «Slogans», «Thought Terminating Cliches» та «Whataboutism» також буде проведено дослідження із застосуванням SMOTE балансування та без нього.

Підхід на основі рекурентних рейромереж включає в себе порівняння 3-х видів архітектур: RNN, LSTM та GRU.

Підхід на основі моделей-трансформерів включає в себе порівняння BERT-подібних моделей: RoBERTa, BERT, ELECTRA.

Для проведення експерименту було створено програмне забезпечення на мові програмування Python, з використанням бібліотек для машинного навчання Sklearn [24], Tensorflow [25], LimeTextExplainer [26], Numpy [27], Pandas [28]. Програмне забезпечення складається із консольного застосунку для навчання моделей машинного навчання, консольного застосунку для виявлення технік пропаганди за маркерами та веб-модуля для візуальної аналітики оцінювання прийнятих результатів обраною моделлю машинного навчання з її оцінками.

Для навчання моделей машинного навчання, що будуть виконувати функції виявлення прийомів пропаганди, буде використано набір даних «emnlp_trans_uk_dataset», що є перекладеним набором даних «emnlp_en_dataset» з відповідністю розмітки на українській мові, взятий з Kaggle-змагань «Disinformation Detection Challenge» [29] з посиланням на «Analysis Project».

Команда «Analysis Project» [30] провела аналіз текстів, виявивши всі фрагменти, які містять пропагандистські прийоми, а також їх тип. Розподіл статей за довжиною у символах наведено на рисунку 3.

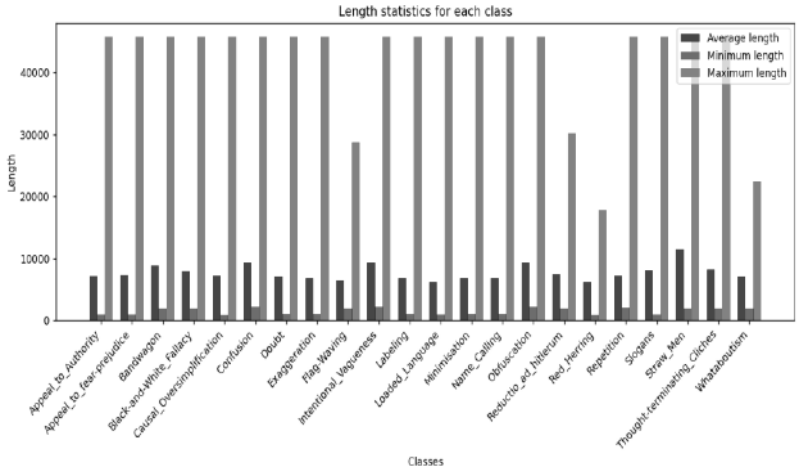

Рисунок 3.*Статистика за довжиною у символах по прийомам пропаганди*

Зокрема, ними створено корпус новинних статей, анотованих вручну на рівні фрагментів за допомогою вісімнадцяти пропагандистських прийомів. Набір даних налічує 788 статей.

Як видно з графіку на рисунку 3, для більшості прийомів пропаганди довжина текстів де вони представлені особливої ролі не грають. Однак, «Flag Waving», «Red Herring», «Reductio ad hitlerum» та «Whataboutism» все ж мають меншу максимальну довжину в текстах, де вони представлені.

Для тренування моделей машинного навчання даний датасет було модифіковано таким чином, щоб текст що містить кожен прийом пропаганди був розміщений в окремому каталозі. Після такого перерозподілу було виведено статистику наявних текстів, що репрезентують прийоми пропаганди. Статистика наведена на рисунку 4. Як видно з рисунку 4 деякі прийоми пропаганди, такі як «Bandwagon», «Confusion», «Intentional Vagueness», «Obfuscation» та «Straw Men» представлені у критично низькій кількості (менше 20 тестів), тому для них окремі класифікатори створені не будуть, ці дані будуть об’єднані у категорію «Інші прийоми пропаганди», однак таким чином, щоб у наявному наборі не були присутні інші прийоми, відмінні від п’яти перерахованих. До прийомів пропаганди, що представлені менш ніж у 100 документах, однак більше ніж 20 буде застосовано SMOTE-балансування під час навчання класифікаторів [31]. До таких категорій належать: «Appeal to

Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» та «Whataboutism».

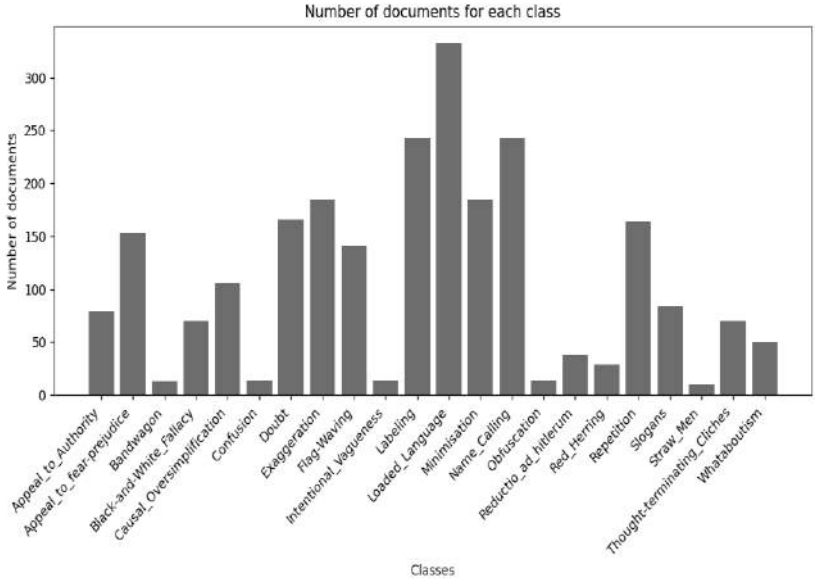

Рисунок 4. *Статистика по кількості текстів, що представляють прийоми пропаганди, шт.*

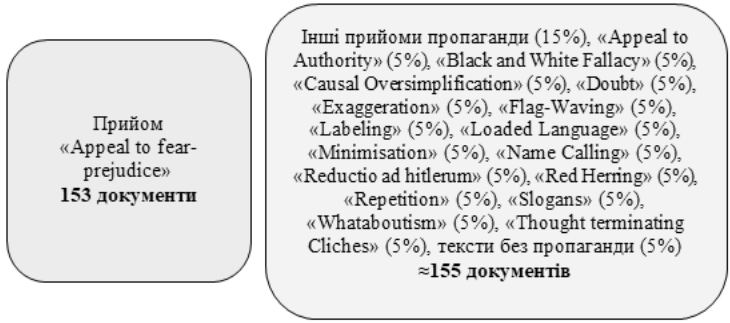

Рисунок 5. *Приклад балансування при формуванні набору даних для навчання та тестування моделі виявлення прийому «Appeal to fear-prejudice»*

Із розглянутого вище набору даних для кожної з 17 типових моделей машинного навчання буде сформовано власний дочірній набір текстів, що буде задовольняти такі вимоги:

- мати тексти з визначеним прийомом пропаганди;

- у противагу використовувати набір «Інші прийоми пропаганди» доповнений текстами без пропаганди та текстами, що представляють інші прийоми пропаганди, відмінні від цільового виду.

Приклад формування набору даних для виявлення прийому «Appeal to fear-prejudice» наведено на рисунку 5.

Отже, у дослідженні буде використано 18 класів: 17 цільових, що є репрезентативними по кількості та відповідають 17 визначеним прийомам пропаганди та 5 об'єднаних в категорію «Інші прийоми пропаганди».

5. Результати та дискусія

Результати дослідження для Traditional machine learning approaches для прийомів «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» та «Whataboutism» без використання SMOTE балансування за метрикою точності наведено в Таблиці 1.

Таблиця 1

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди до SMOTE балансування за метрикою Accuracy

Techniques of propaganda	Regression	SVM	Random Forest	Naive Bayes
Appeal to Authority	0.57	0.63	0.55	**0.64**
Black and White Fallacy	0.55	**0.64**	0.51	0.58
Reductio ad hitlerum	**0.68**	0.56	0.61	0.59
Red Herring	**0.61**	**0.61**	0.59	**0.61**
Slogans	0.62	**0.63**	0.56	0.62
Thought terminating Cliches	0.59	0.58	**0.63**	0.58
Whataboutism	0.62	**0.65**	0.59	0.57

Як видно з таблиці 1, точність виявлення прийомів пропаганди коливається від 0.51 до 0.68, що є доволі низьким показником.

Наступним етапом для даних прийомів пропаганди було застосовано SMOTE балансування, збільшивши таким чином кількість навчальних зразків до рівня не менше 100. Результат експерименту наведено у таблиці 2.

Як видно з таблиці 2, застосування SMOTE балансування дало позитивні результати для більшості прийомів пропаганди, однак для «Slogans» результат покращення не дав.

Це пов'язано з тим, що кількість навчальних зразків близька до граничної і є достатньою для навчання запропонованих версій машинного навчання.

Таблиця 2

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди зі SMOTE балансування за метрикою Accuracy

Techniques of propaganda	Regression	SVM	Random Forest	Naive Bayes
Appeal to Authority	0.59	**0.67**	0.54	0.60
Black and White Fallacy	0.61	0.62	0.55	**0.64**
Reductio ad hitlerum	**0.69**	0.63	0.62	0.58
Red Herring	**0.69**	0.64	0.58	0.61
Slogans	0.62	**0.63**	0.56	0.62
Thought terminating Cliches	0.62	**0.68**	0.62	0.61
Whataboutism	0.64	**0.66**	0.58	0.56

Для решти прийомів SMOTE балансування не застосовувалось, та результати застосування традиційного підходу машинного навчання наведено у таблиці 3.

Таблиця 3

Традиційний підхід на основі машинного навчання для виявлення прийомів пропаганди без SMOTE балансування за метрикою Accuracy

Techniques of propaganda	Regression	SVM	Random Forest	Naive Bayes
Appeal to fear-prejudice	**0.62**	0.61	0.55	0.59
Causal Oversimplification	0.58	**0.62**	0.59	0.55
Doubt	0.6	0.58	**0.63**	0.59
Exaggeration	0.61	**0.63**	0.61	0.53
Flag-Waving	0.62	0.61	**0.64**	0.6
Labeling	**0.67**	0.62	0.61	0.61
Loaded Language	**0.6**	0.59	**0.6**	0.58
Minimisation	**0.62**	0.61	0.61	0.60
Name Calling	0.55	0.59	0.57	**0.61**
Repetition	0.69	**0.7**	0.61	0.55

Результати з таблиці 3 також коливаються від 0.6 до 0.67, що має схожий результат із застосуванням SMOTE балансування (таблиця 2). Однак отримані результати також не є задовільними. Ілюстрація експерименту щодо застосування традиційного підходу машинного навчання наведена на рисунку 6.

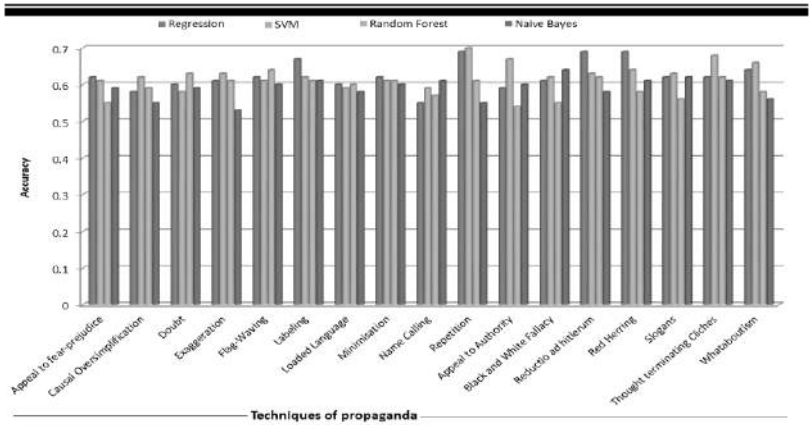

Рисунок 6. *Порівняння точності моделей для традиційного підходу машинного навчання*

Таблиця 4

Підхід на основі рекурентних нейромереж для виявлення прийомів пропаганди за метрикою Accuracy

Techniques of propaganda	RNN	LSTM	GRU
Appeal to fear-prejudice	0.69	0.69	**0.71**
Causal Oversimplification	0.73	0.71	**0.76**
Doubt	**0.75**	0.7	0.74
Exaggeration	0.64	0.72	**0.75**
Flag-Waving	0.69	0.7	**0.79**
Labeling	0.69	0.73	**0.8**
Loaded Language	**0.71**	0.7	0.68
Minimisation	**0.78**	**0.78**	0.74
Name Calling	**0.76**	0.74	**0.76**
Repetition	0.74	0.75	**0.76**
Appeal to Authority	0.71	0.72	**0.73**
Black and White Fallacy	0.7	0.68	**0.72**
Reductio ad hitlerum	**0.75**	0.68	0.71
Red Herring	0.65	**0.72**	0.70
Slogans	0.74	0.68	**0.75**
Thought terminating Cliches	0.63	**0.66**	0.65
Whataboutism	0.67	**0.69**	**0.69**

Наступним експериментом проводилось дослідження застосування An approach based on recurrent rail networks, що включало в себе порівняння використань 3-х видів архітектур: RNN, LSTM та GRU. Дані експерименту без використання SMOTE балансування наведено у таблиці 4.

Як видно з даних таблиці 4, результати для усіх прийомів пропаганди окрім Thought terminating Cliches, є вищими та знаходяться у діапазоні від 0.66 до 0.8. Однак до даного прийому пропаганди буде в подальшому застосовано SMOTE балансування, що можливо дозволить покращити показник. Наступним експериментом буде використання SMOTE балансування до навчання нейромережевих моделей для «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches» та «Whataboutism». Із таблиці 5 видно, що SMOTE балансування дає позитивний ефект на точність виявлення прийомів пропаганди. Не вдалося покращити виявлення прийому «Reductio ad hitlerum», де результати до SMOTE балансування були на 0.01 вище, а також «Appeal to Authority» та «Black and White Fallacy» залишились на тому ж рівні, що і до виконання балансування. На рисунку 7 показано порівняння найвищих отриманих показників точності для рекурентних нейромережевих моделей.

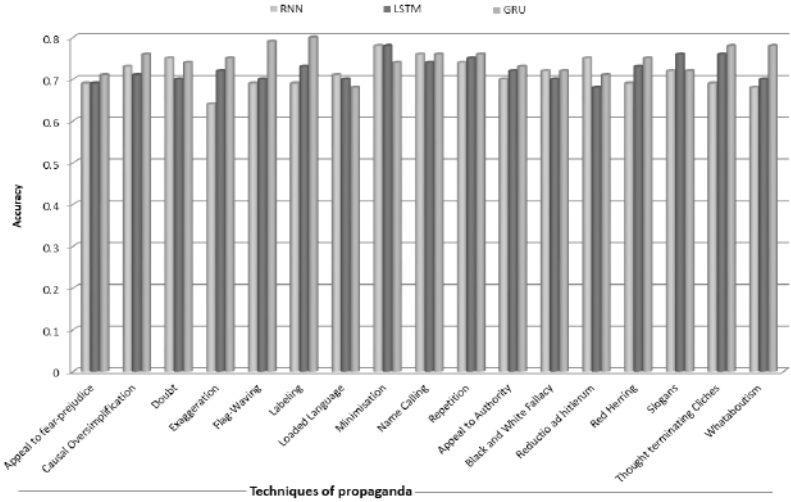

Рисунок 7. *Порівняння точності моделей для підходу на основі рекурентних нейронних мереж*

Останнім етапом дослідження є використання підходу на основі використання моделей-трансформерів, що включає в себе порівняння BERT-подібних моделей: RoBERTa, BERT, ELECTRA. Були використані попередньо натреновані моделі з ресурсу Hugging Face [32], які було донавчені вищеописаним способом протягом 3-х епох навчання. Отримані результати без використання SMOTE балансування наведені у таблиці 6.

Таблиця 5

Підхід на основі рекурентних нейромереж для виявлення прийомів пропаганди за метрикою Accuracy із SMOTE балансуванням

Techniques of propaganda	RNN	LSTM	GRU
Appeal to Authority	0.7	0.72	**0.73**
Black and White Fallacy	**0.72**	0.7	**0.72**
Reductio ad hitlerum	0.73	**0.74**	**0.74**
Red Herring	0.69	0.73	**0.75**
Slogans	0.72	**0.76**	0.72
Thought terminating Cliches	0.69	0.76	**0.78**
Whataboutism	0.68	0.7	**0.78**

Таблиця 6

Підхід на основі нейромереж-трансформерів для виявлення прийомів пропаганди за метрикою Accuracy

Techniques of propaganda	bert-base-multilingual-cased	roberta-base	ukr-electra-base
Appeal to fear-prejudice	0.81	0.8	**0.87**
Causal Oversimplification	0.78	0.79	**0.82**
Doubt	**0.93**	0.9	0.87
Exaggeration	**0.8**	**0.8**	**0.8**
Flag-Waving	**0.92**	0.9	0.89
Labeling	**0.96**	0.94	**0.96**
Loaded Language	0.93	**0.97**	0.94
Minimisation	0.89	0.86	**0.9**
Name Calling	**0.92**	**0.92**	0.91
Repetition	0.93	**0.94**	**0.94**
Appeal to Authority	0.87	**0.89**	0.88
Black and White Fallacy	0.89	**0.91**	0.88
Reductio ad hitlerum	0.85	**0.87**	0.86
Red Herring	0.67	**0.8**	0.78
Slogans	0.84	**0.86**	0.83
Thought terminating Cliches	**0.8**	0.73	0.79
Whataboutism	**0.79**	0.78	0.78

Із даних таблиці 6 видно, що BERT-подібні нейромережеві архітектури значно краще виявляють прийоми пропаганди, порівняно з рекурентними та традиційним підходами до навчання. Це пояснюється тим, що такі архітектури

є контекстно-орієнтованими, що є важливим аспектом для виявлення прийомів пропаганди. Застосування SMOTE балансування дозволило підвищити точність виявлення «Red Herring» нейромережевою моделлю ukr-electra-base до 0.89, а також «Whataboutism» до 0.83 з використанням bert-base-multilingual-cased. Порівняння найвищих оцінок за метрикою точності для 3-х розглянутих підходів наведено на рисунку 8.

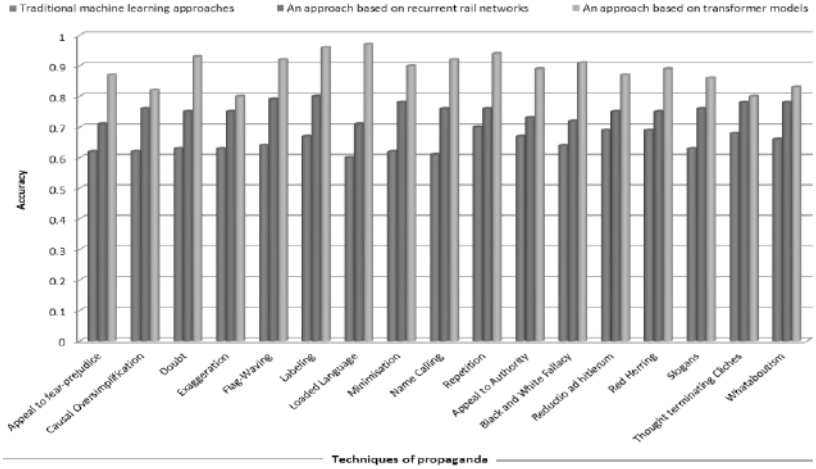

Рисунок 8. *Порівняння точності моделей альтернативних підходів для виявлення технік пропаганди*

Як видно з рисунку 8, традиційний підхід на основі машинного навчання очікувано показав гірші результати, оскільки він не вміє бачити контекст, що є важливим для виявлення технік пропаганди. Рекурентні нейромережеві моделі хоч і показали результати вищі від традиційного підходу, однак все ще мають проблеми з обробкою довгих залежностей. Найвищі результати з проведеного експерименту виявились у підході на основі моделей-трансформерів, це пояснюється використовуваними механізмами самоуваги, які дозволяє кожному елементу послідовності безпосередньо взаємодіяти з усіма іншими елементами. Це дозволяє ефективно захоплювати довготривалі залежності, що характерно для проявів технік пропаганди. Отримані результати забезпечили виявлення різних пропагандистських прийомів з мінімальною точністю 79,03% (мінімальні значення точності отримані для методики «Whataboutism»), що краще за відомі аналоги [8] щодо виявлення пропаганди незалежно від використовуваних методик. Порівняно з відомими аналогами [7] підвищилась точність виявлення різних пропагандистських прийомів:

– для техніки «Appeal to Authority», точність виявлення зросла на 9.81% (існуючий метод 77.27%, розроблений метод 87.08%);

– для техніки «Causal Oversimplification», точність виявлення зросла на 11.99% (існуючий метод 70.1%, розроблений метод 82.09%);

– для техніки «Doubt», точність виявлення зросла на 75.32% (існуючий метод 17.78%, розроблений метод 93.1%);

– для техніки «Exaggeration», точність виявлення зросла на 26.04% (існуючий метод 54.17%, розроблений метод 80.21%);

– для техніки «Flag-Waving», точність виявлення зросла на 27.65% (існуючий метод 64.52%, розроблений метод 92.17%);

– для техніки «Labeling», точність виявлення зросла на 48.57% (існуючий метод 47.43%, розроблений метод 96.0%);

– для техніки «Loaded Language», точність виявлення зросла на 42.9% (існуючий метод 54.17%, розроблений метод 97.07%);

– для техніки «Name Calling», точність виявлення зросла на 44.6% (існуючий метод 47.43%, розроблений метод 92.03%);

– для техніки «Repetition», точність виявлення зросла на 58.11% (існуючий метод 35.98%, розроблений метод 94.09%);

– для техніки «Appeal to Authority», точність виявлення зросла на 11.84% (існуючий метод 77.18%, розроблений метод 89.02%);

– для техніки «Black and White Fallacy», точність виявлення зросла на 36.68% (існуючий метод 54.55%, developed method 91.23%);

– для техніки «Reductio ad hitlerum», точність виявлення зросла на 62.31% (існуючий метод 25.0%, розроблений метод 87.31%);

– для техніки «Red Herring», точність виявлення зросла на 41.07% (існуючий метод 39.22%, розроблений метод 80.29%);

– для техніки «Slogans», точність виявлення зросла на 10.54% (існуючий метод 75.5%, розроблений метод 86.04%);

– для техніки «Thought terminating Cliches», точність виявлення зросла на 26.74% (існуючий метод 53.57%, розроблений метод 80.31%);

– для техніки «Whataboutism», точність виявлення зросла на 39.81% (існуючий метод 39.22%, розроблений метод 79.03%).

Приклад візуальної поясненності [33, 34] щодо виявлення прийому пропаганди «Repetition» наведено на рисунку 9 (використано оригінальний текст повсякденною українською мовою із збереженням орфографії та помилок). Як видно з рисунку 9, є багаторазові повторення фраз на кшталт «економічні мігранти», «мусульманський», «Орбан» того. З означення виду пропаганди «Repetition», це є «повторення того самого повідомлення знову і знову, щоб глядачі зрештою прийняти це». Отже, запропонований метод дозволяє ефективно виявляти прийоми пропаганди, та має перевагу у точності в порівнянні із запропонованими моделями що використовують підхід багатокласової класифікації.

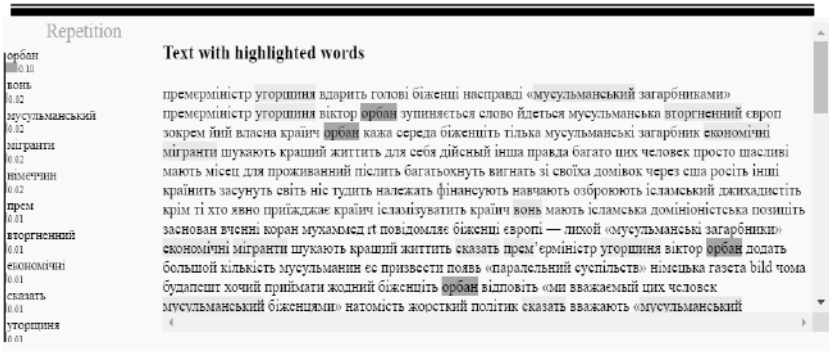

Рисунок 9. *Візуальна аналітика щодо виявлення прийому пропаганди «Repetition» розробленим програмним забезпеченням*

Експерименти, представлені в статті, проводилися з використанням різних можливостей бібліотеки SKLearn. У цій роботі представлені максимальні результати, яких вдалося досягти авторам емпіричним шляхом. Питання конфігурації та вибору гіперпараметрів є окремою проблемою, яка виходить за рамки питань, що розглядаються.

6. Висновки

Створено метод виявлення та класифікації технік пропаганди у текстовому контенті засобами штучного інтелекту, що дозволяє перетворювати вхідні дані у вигляді тексту для аналізу та навчених моделей машинного навчання у вихідні дані, які містять числові оцінки наявності кожного з прийомів пропаганди та розмічений текст з візуальною аналітикою присутності детектованих маркерів пропаганди. Було проведено дослідження, що дозволяє визначати 17 основних прийомів пропаганди, таких як: «Appeal to fear-prejudice», «Causal Oversimplification», «Doubt», «Exaggeration», «Flag-Waving», «Labeling», «Loaded Language», «Minimisation», «Name Calling», «Repetition», «Appeal to Authority», «Black and White Fallacy», «Reductio ad hitlerum», «Red Herring», «Slogans», «Thought terminating Cliches», «Whataboutism».

У рамках дослідження було порівняно 3 найчастіше застосовувані підходи: традиційний підхід машинного навчання, підхід на основі рекурентних нейронних мереж та підхід на основі трансформерних моделей. Традиційні підходи машинного навчання очікувано показали гірші результати, оскільки вони не здатні враховувати контекст, що є важливим для виявлення технік пропаганди. Досягнута точність для традиційного підходу становила від 0.60 до 0.67. Рекурентні нейромережі, хоча й перевершили традиційні підходи, все ж мають труднощі з обробкою довгих залежностей. Для цього підходу було досягнуто точності від 0.66 до 0.80. Найвищі результати були досягнуті підходом на основі трансформерних моделей, завдяки використанню механізмів самоуваги, які дозволяють кожному елементу послідовності безпосередньо взаємодіяти з усіма іншими елементами. Це забезпечує

ефективне захоплення довготривалих залежностей, що є характерним для технік пропаганди. Даний підхід дозволив виявляти техніки пропаганди з точністю до 0.96. Отримані результати забезпечили виявлення різних пропагандистських прийомів з мінімальною точністю 79,03% (мінімальні значення точності отримані для методики «Whataboutism»), що краще за відомі аналоги [8] щодо виявлення пропаганди незалежно від використовуваних методик. Порівняно з відомими аналогами [7] покращилась точність виявлення різноманітних прийомів пропаганди: за методикою «Appeal to Authority» точність виявлення зросла на 9,81%, за методикою «Причинне спрощення» — на 11,99%, за методикою «Causal Oversimplification» точність виявлення зросла на 75,32% для методики «Doubt», точність виявлення зросла на 26,04% для техніки «Exaggeration», точність виявлення зросла на 27,65% для техніки «Flag-Waving», точність виявлення зросла на 48,57% для «Labeling» методики, точність виявлення зросла на 42,9% для методики «Loaded Language», точність виявлення зросла на 44,6% для техніки «Name Calling», точність виявлення зросла на 58,11% для техніки «Repetition», точність виявлення зросла на 11,84% для техніки «Appeal to Authority», точність виявлення зросла на 36,68% для техніки «Black and White Fallacy», точність виявлення зросла на 62,31% для техніки «Reductio ad hitlerum», точність виявлення зросла на 41,07% для «Red Herring», точність виявлення зросла на 10,54% для техніки «Slogans», точність виявлення зросла на 26,74% для техніки «Thought terminating Cliches», точність виявлення зросла на 39,81% для техніки «Whataboutism».

Подальші дослідження будуть спрямовані на розширення датасету для навчання та пошуку додаткових міток у текстах, що характеризують прийоми пропаганди, таких як наявність булінгу, емоційна тональність тощо, що дозволить зробити більш поясненим рішення моделі машинного навчання та дозволить зробити більш точним виявлення прийомів.

7. Література

[1] A. Horak, R. Sabol, O. Herman, V. Baisa, Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis, Expert Systems with Applications, Volume 251, 202. doi:10.1016/j.eswa.2024.124085.

[2] A. Bhattacharjee, H. Liu, Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text? SIGKDD Explor, Newsl, 25, pp. 14–21, 2023. doi:10.1145/3655103.3655106.

[3] G. Faye, B. Icard, M. Casanova, J. Chanson, F. Maine, F. Bancilhon, G. Gadek, G. Gravier, P. Egre, Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification, in: Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language, pp. 62–72, Malta, Association for Computational Linguistics.

[4] P. Vijayaraghavan, S. Vosoughi, TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for

Computational Linguistics, pp. 3433–3448, Seattle, United States, Association for Computational Linguistics.

[5] M. Abdullah, O. Altiti, R. Obiedat, Detecting Propaganda Techniques in English News Articles using Pre-trained Transformers, in: 2022 13th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2022, pp. 301-308. doi: 10.1109/ICICS55353.2022.9811117.

[6] D.G. Jones, Detecting Propaganda in News Articles Using Large Language Models, Eng OA, 2(1), 2024, pp. 1-12. URL: https://www.opastpublishers.com/peer-review/detecting-propaganda-in-news-articles-using-large-language-models-6952.html.

[7] G. D. S. Martino, A. Barron-Cedeno, H. Wachsmuth, R. Petrov, P.Nakov, SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1377–1414, Barcelona (online), International Committee for Computational Linguistics.

[8] G. Martino, S. Yu, A. Barron-Cedeno, R. Petrov, P. Nakov, Fine-Grained Analysis of Propaganda in News Article, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5640-5650, 2019. doi:10.18653/v1/D19-1565.

[9] Analytics Vidhya, Guide on Support Vector Machine (SVM) Algorithm, 2024. URL: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

[10] S. Mann, D. Yadav, D.Rathee, Identification of Racial Propaganda in Tweets Using Sentimental Analysis Models: A Comparative Study, in: Swaroop, A., Kansal, V., Fortino, G., Hassanien, A.E. (eds) Proceedings of Fourth Doctoral Symposium on Computational Intelligence. DoSCI 2023. Lecture Notes in Networks and Systems, vol 726, Springer, Singapore. doi: 0.1007/978-981-99-3716-5_28.

[11] L. Syed, A. Alsaeedi, L. A. Alhuri, H. R. Aljohani, Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda, Array, Volume 19, 2023. doi:10.1016/j.array.2023.100309.

[12] Krak I., Zalutska O., Molchanova M., Mazurets O., Bahrii R., Sobko O., Barmak O. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 16-28. URL: https://ceur-ws.org/Vol-3688/paper2.pdf.

[13] Geeks for geeks, What is LSTM – Long Short Term Memory, 2024. URL: https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/.

[14] A.M.U.D. Khanday, Q.R. Khan, S.T. Rabani, M.A. Wani, M. ELAffendi, Propaganda Identification on Twitter Platform During COVID-19 Pandemic Using LSTM, in: Abd El-Latif, A.A., Maleh, Y., Mazurczyk, W., ELAffendi, M., I. Alkanhal, M. (eds) Advances in Cybersecurity, Cybercrimes, and Smart Emerging

Technologies, CCSET 2022, Engineering Cyber-Physical Systems and Critical Infrastructures, vol 4. Springer, Cham. doi:10.1007/978-3-031-21101-0_24.

[15] Dev, Large Language Models: Comparing Gen 1 Models (GPT, BERT, T5 and More), 2024. URL: https://dev.to/admantium/large-language-models-comparing-gen-1-models-gpt-bert-t5-and-more-74h.

[16] Hugging Face, BERT, 2024. URL: https://huggingface.co/docs/transformers/model_doc/bert.

[17] Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 344-356. URL: https://ceur-ws.org/Vol-3387/paper26.pdf

[18] Hugging Face, DistilBERT, 2024. URL: https://huggingface.co/distilbert/distilbert-base-uncased.

[19] D. Chaudhari, A. V. Pawar, Empowering Propaganda Detection in Resource-Restraint Languages: A Transformer-Based Framework for Classifying Hindi News Articles, Big Data and Cognitive Computing, 2023, 7(4):175. doi:10.3390/bdcc7040175.

[20] A. Malak, D. Abujaber, A. Al-Qarqaz, R. Abbott, M. Hadzikadic, Combating propaganda texts using transfer learning, IAES International Journal of Artificial Intelligence (IJ-AI). Volume 12, pp. 956-965, 2023. doi: 10.11591/ijai.v12.i2.pp956-965

[21] A. Malak, D. Abujaber, A. Al-Qarqaz, R. Abbott, M. Hadzikadic, Combating propaganda texts using transfer learning, IAES International Journal of Artificial Intelligence (IJ-AI). Volume 12, pp. 956-965, 2023. doi: 10.11591/ijai.v12.i2.pp956-965.

[22] I. Krak, O. Barmak, O. Mazurets, The practice investigation of the information technology efficiency for automated definition of terms in the semantic content of educational materials. CEUR Workshop Proceedings, 2016, vol.1631, pp. 237–245. doi:10.15407/pp2016.02-03.237.

[23] C3.ai, What is Local Interpretable Model-Agnostic Explanations (LIME)?, 2024. URL: https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/.

[24] Scikit-learn, Machine Learning in Python, 2024. URL: https://scikit-learn.org/stable/.

[25] Tensorflow, An end-to-end platform for machine learning. Machine Learning in Python, 2024. URL: https://www.tensorflow.org.

[26] GitHub, Lime, 2024. URL: https://github.com/marcotcr/lime.

[27] Numpy, The fundamental package for scientific computing with Python, 2024. URL: https://numpy.org.

[28] Pandas, Pandas, 2024. URL: https://pandas.pydata.org.

[29] Kaggle, Disinformation Detection Challenge, 2024. URL: https://www.kaggle.com/competitions/disinformation-detection-challenge/data?select=emnlp_trans_uk_dataset.
[30] Propaganda, Propaganda Analysis Project, 2024. URL: https://propaganda.qcri.org/index.html.
[31] Y. Elor, H. Averbuch-Elor, To SMOTE, or not to SMOTE. URL: https://arxiv.org/abs/2201.08528.
[32] Hugging Face, The AI community building the future, 2024. URL: https://huggingface.co/.
[33] V. Slobodzian, M. Molchanova, O. Kovalchuk, O. Sobko, O. Mazurets, O. Barmak, I. Krak. An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. 2022 12th International Conference on Advanced Computer Information Technologies, ACIT 2022. 2022. pp. 400-405. doi: 10.1109/ACIT54803.2022.9913162
[34] O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets. Book Chapter. Lecture Notes on Data Engineering and Communications Technologies. 2023. Vol. 149. pp. 591–607. doi: 10.1007/978-3-031-16203-9_33

METHOD FOR DETECTING AND CLASSIFYING PROPAGANDA TECHNIQUES IN TEXTUAL CONTENT USING ARTIFICIAL INTELLIGENCE

M. Molchanova ORCID: 0000-0001-9810-936X
Khmelnytskyi National University, Ukraine
E-mail: m.o.molchanova@gmail.com

Abstract. *The paper is devoted to the creation and approbation of the method for neural network detecting propaganda techniques by markers with visual analytic, which allows converting input data in the form of text for analysis and supervised machine learning models into output data containing numerical estimates of the presence of each propaganda technique and marked-up text with visual analytical presence of detected propaganda markers. Research was conducted that allows us to detect 17 main propaganda techniques. The study compared the 3 most commonly used approaches: a traditional machine learning approach, an approach based on recurrent neural networks, and an approach based on transformer models. The highest results were achieved by the transformer model approach, which uses self-attention mechanisms that allow each element of the sequence to interact directly with all other elements. This ensures efficient capture of long-term dependencies, which is typical for propaganda techniques. This approach allowed us to detect propaganda techniques with an accuracy of 0.96.*

Keywords: *BERT, RNN, propaganda techniques, detecting propaganda, propaganda markers, visual analytics*